Does responsibility encompass ethicality and accountability in language assessment?

Albert Weideman*

University of the Free State, South Africa

ARTICLE INFO

ABSTRACT

A central conceptual issue in language assessment in general, and in the work of Alan Davies in particular, is never fully resolved. How is responsible language test design related to ethicality? This unfinished business goes back to the unresolved debate about validity and validation, that has resulted in a loss of conceptual clarity about sound language assessment. The contradictions inherent in first declaring validity to be the overarching concept, and then attempting to expand it by promoting other ideas to be the prime considerations, further confuse the issue. This contribution argues that such expansion is unhelpful. A way out is to look at language test design as being responsive to certain typical and general conditions. In that relationship, between the (subjective) making of the test as an artefact that presents us with a response to certain design principles, and the designed object, the actual test itself, lies a potential way out of the impasse.

© 2016 Elsevier Ltd. All rights reserved.

1. Diligence, quality and responsible test design

Though this contribution will deal with how some ideas and principles of language test design conventionally considered to be important are conceptualized, it is written primarily from the perspective of a language test developer. Language assessment design presents an onerous set of responsibilities for applied linguists who work in that subfield, and few would argue with the premise that the design of language tests has to be done with the greatest deliberation, diligence and care. Language tests have to be designed and developed in a way that is worthy of professional conduct (Davies, 1997). Responsible test designers thus regard their work as a profession; it is not surprising that Davies and Elder (2005: 800) observe: "What professions exist to do is to combine field expertise with a proper moral and social concern to act responsibly in normal settings." The combination they refer to here is the theme of this paper: the relation between the technical know-how to make a language test, and its sometimes unarticulated and at times even obscure connection with other dimensions, concepts, and ideas that inform test design and the subsequent use of tests.

The quality of the assessments designed by language testers becomes prominent where tests of language ability are used for medium to high stakes purposes, and acutely so when they are administered on scale. Many high stakes tests one finds at the hinge-points of education or access to employment opportunity, for example just before the end of pre-tertiary education, or before entry into higher education or the world of work (cf. too Joseph, 2016). Especially where secondary school exit examinations are administered as public, government-initiated tests, they are almost without fail high stakes assessments. Their results will inevitably be used as evidence of potential employability in the world of work, or for admission to tertiary

---

* Office of the Dean, Humanities, University of the Free State, Nelson Mandela Avenue, Park West, 9301 Bloemfontein, South Africa.
E-mail address: albert.weideman@ufs.ac.za.

education. What increases their impact even more is that they are often nationally organized – and in the case of some large scale, commercial language assessments, internationally. In all such cases, the quality of the assessment on a national or global scale further significantly increases their potentially beneficial or disadvantageous impact. It is no coincidence then that their worth and quality should be defensible both publicly and among experts.

There is a consistent thread in the work of Alan Davies and those who collaborated with him on all of these themes and issues. This contribution will analyse some of the most significant discussions in the work of Davies *c.s.* from the perspective of those applied linguists who are engaged in the design and development of language tests. It will ask what the enduring meaning of these discussions might be for the design of language assessments. It is noteworthy, in the first instance, that Davies not only wrote on language assessment as a subfield of applied linguistics (for this, cf. McNamara and Roever, 2006: 255; cf. too McNamara, 2006; Weideman, 2006), but in such discussion consistently brought accountability for such applied linguistic designs into play. Referring to the three prime applied linguistic artefacts (Weideman, 2014, 2017), namely language courses, language tests and language policies, Davies (2008: 298) states that "applied linguistics is prepared in its curricula and its assessments and in its planning… to be accountable", adding that it accomplishes that, amongst other things, "by theorising practice".

How does that accountability fit in with, or relate to responsible test design, however? How do all the various concepts and ideas, such as transparency, accessibility, fairness, ethical codes (Boyd and Davies, 2002), and standards (Davies, 1997, 2008) cohere, if at all? Does accountability depend purely and solely on predetermined or professional standards? Or does it perhaps depend even more on locally relevant and contextualized codes of practice? What are the necessary, and what the sufficient safeguards for fairness in language testing? This contribution will offer a possible alternative perspective to the ideas currently in the mainstream, one that is potentially more integrated and possibly clearer conceptually. It will do so by proposing that the idea of responsible test design offers a coherent framework of principles that encompasses the notions of transparency, utility, accessibility and accountability – in fact the ethicality of assessment generally. It will proceed, like Davies, from the assumption that language test design falls within the domain of constructing applied linguistic artefacts. It may also provide further insight into the role of standards in achieving responsible design.

2. Unresolved conceptual issues

In discussing how standards may or should affect language assessment, Davies (2008) deals specifically with standards as a set of criteria for assessing the appropriateness of the language tests that are designed to measure language performance. Standards as procedures and yardsticks for test design are thus his prime concern, as well as that of this discussion. Recognizing that reliability, validity, appropriateness, utility, accessibility, and theoretical defensibility are necessary pre-requisites for tests, Davies (2008: 491f.) also emphasizes that these may not be sufficient: a public demonstration of their worth that clarifies their political, ethical and juridical dimensions is also required. The idea of consequential validity (Messick, 1988, 1989: 20, 88; Davies, 1997, 2011: 335; Davies and Elder, 2005: 798; McNamara and Roever, 2006; Weideman, 2012; Boyd and Davies, 2002: 304, 306), of gauging the social, political and economic impact of language tests, clearly is what Davies (2008) has in mind. Perhaps, if one considers his discussion and analysis as a whole, that discussion may even be shown to encompass the broader notion of the appropriateness of the public provision of language instruction and the assessment at school of such language instruction. Elsewhere, that is sometimes referred to as the contextual or ecological validity of a course or an assessment (Arzubiaga et al., 2008).

At the same time, we should note that this discussion (Davies, 2008) refers to and builds upon earlier ones in Davies's work, especially his involvement in helping to draft, support, and promote the Code of Ethics of the International Language Testing Association (ILTA) (Davies, 1997: 336). In these analyses, the weakness of the language testing profession to enforce adherence to a set of standards is noted (Boyd and Davies, 2002: 307), but also that the cause is not lost: "For the emerging profession of language testing it is not too late to build in openness to its professional life" (Boyd and Davies, 2002: 312). As we can see, in Davies's perspective the transparency of professional practice remains a powerful tenet of his arguments about accountability. Thus openness, a readiness by the professional to provide "clearer public information about the professional expertise needed for language test construction" (Davies, 1997: 338), is introduced as a means of language test designers being publicly held to their own professional standards. Openness or transparency becomes the basis for accountability, the other closely related, significant theme in Davies's analyses: as can be expected, Boyd and Davies (2002) explicitly take accountability as their central point.

These discussions about 'standards' therefore immediately become related to a whole range of principles for evaluating the quality of language tests. That range might include all the familiar issues: transparency, accessibility, utility, accountability, fairness, care, integrity, trust, humility and the like. These are in fact the main threads, too, of the ILTA Code of Ethics that Boyd and Davies (2002) discuss as the proposed backbone of professional conduct for language testers. Yet the conceptual question that is neither asked nor answered when bringing them into such a code or its subsequent discussion is whether they may be treated as disparate issues (individually articulated 'principles', with annotations, as in the Code), or whether (more plausibly) they should also be treated as related, and if the latter, what that relationship is. The problematic nature of the relationship becomes evident when we come across statements such as: "While I can accept that ethics in language testing does include validity, whether it is wider in scope remains an unresolved question" (Davies, 1997: 335). Or consider this claim, again with reference to the ILTA Code of Ethics: "… what we are seeing in the professionalizing and ethicalizing of language testing is a wider and wider understanding of validity" (Davies, 2008: 491).